BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37276
doi: 10.3897/biss.3.37276

OPEN ACCESS

Conference Abstract

# Big Data Knowledge of Major Lineages of Life and Priorities for Genomic Research

Jonathan Coddington[‡], Katharine Barker[‡], Gabriele Droege[§], Ole Seberg[|]

‡ National Museum of Natural History, Smithsonian Institution, Washington, D.C., United States of America
§ Botanic Garden and Botanical Museum Berlin, Berlin, Germany
| Natural History Museum Denmark, Copenhagen, Denmark

Corresponding author: Jonathan Coddington (coddington@si.edu)

Received: 13 Jun 2019 | Published: 19 Jun 2019

Citation: Coddington J, Barker K, Droege G, Seberg O (2019) Big Data Knowledge of Major Lineages of Life and Priorities for Genomic Research. Biodiversity Information Science and Standards 3: e37276.
https://doi.org/10.3897/biss.3.37276

## Abstract

Genomic science is revolutionizing and accelerating biodiversity research. For collections-based institutions to continue to lead and support biodiversity research, they must adapt to this new reality. Simultaneously, “big data” is accumulating so rapidly that we have unprecedented capacity to plan strategically to use genomics to advance basic and applied science on multiple fronts. For example, seven “big data” sources have the following numbers of records (2018 data): Global Biodiversity Information Facility (GBIF), ~1B; Biodiversity Heritage Library (BHL), ~3.6M; National Center for Biotechnology Information (NCBI), ~220M; Open Tree of Life (OToL), 1.9M; Barcode of Life Data System (BOLD), ~6.3M; Encyclopedia of Life (EOL), ~99K; Global Genome Biodiversity Network (GGBN), ~2M. Collectively, they offer more than 1.2B records on biodiversity. At the scale of species (~2M described, multiple millions undescribed), these data are still too sparse to permit comprehensive conclusions. At the scale of families (i.e. deeper clades of life), the situation is far more promising: about 9,911 families are known, and relatively few are discovered each year. This suggests that at the family rank (and above), our knowledge of life on Earth is reasonably complete. Approximately 160,000 valid and accepted genera exist, but certainly many new genera await discovery and description. Genomics is the fastest way to group species into more inclusive lineages such as genera and families, and is certainly faster than traditional alpha taxonomy. Synergistically, these “big data” answer four

important questions at deeper clade levels: What is it? Where is it? What do we know about it? What do we know about its genome? Approximately 4,500 eukaryotic genomes have been sequenced. The converse of what we know is what we do not know, another meaning of “dark taxa.” We can use the distribution and density of big data at deeper clade levels (families, genera) to quantitatively analyze “dark taxa” and therefore to strategically optimize knowledge and preservation of biodiversity at a global scale. Technicalities of the quantitative prioritization scheme are debatable, but some initial, simple scoring systems can help to prioritize lineages for collection and genetic research so as to most efficiently illuminate regions in the tree of life that that are neither preserved, imaged, geo-located, studied, nor known genomically. This analysis presents criteria and goals for collaborating to build a global genomic collection to maximize efficient acquisition of biodiversity genomic knowledge, and identifies the most valuable and highest priority taxa for genomic research.

## Keywords

Families of Life, barcode, GBIF, EOL, NCBI, BOLD, GGBN, phylogeny, phylogenetic diversity, genome

## Presenting author

Jonathan A. Coddington

## Presented at

Biodiversity_Next 2019